# Reinventing Environmental Information (REI)
# Interim Chemical Identification Data Standard

## I. Introduction

At the Common Sense Initiative (CSI) Council Meeting held on July 21, 1997, Administrator Carol Browner and Deputy Administrator Fred Hansen announced that the Environmental Protection Agency (EPA) would pursue three important information management (IM) reforms:

(1) Establishing key data standards to improve the value of environmental information and enabling data sharing and integration;
(2) Providing universal voluntary access to electronic reporting to reduce burden and improve data quality and timeliness; and
(3) Implementing these data standards and electronic reporting reforms in the Agency's national systems in partnership with the States through the One Stop program.

Together, these reforms provide the basis for REI: "Reinventing Environmental Information." REI focuses on incorporating data standards and electronic reporting into EPA's national systems, with priority placed on the Agency's compliance systems. In addition, EPA will enhance its information management processes to ensure these efforts are successful. Milestones for data standards include:

- Promulgating interim standards for six priority data standards;
- Developing business rules/processes for implementing the standards and promulgating final standards;
- Establishing a central Agency program to support implementation of the standards by EPA and the States; and
- Implementing data standards and business practices in national systems and accepting new data in the standard format from all participating States.

**PRIORITY DATA STANDARDS**

Year 2000 (Y2K) Date
Facility Identification
Standard Industrial Classification (SIC) Code
Latitude/Longitude
Biological Taxonomy
Chemical Identification

The goal of this document is to present a standard for the identification of chemical substances and to make recommendations for its implementation. This standard is intended to:

- Provide a common and consistent way to identify and represent chemical substances across the Agency;
- Provide EPA with a set of common names for the chemical substances it regulates or for which it is responsible;
- Provide a way to reference data about chemical substances across EPA systems and provide a basis for searching for chemicals in these systems in an automated way; and
- Provide a consistent representation for all types of chemical substances in which the Agency has an interest, including single fully-defined chemicals, known, variable and unknown composition chemicals, regulatory classes, regulatory categories and CBI protected chemicals (Table 1 Appendix).

## II. The Chemical Identification Standard

This Standard specifies the key data elements necessary to constitute a viable chemical substance record. This standard requires that a complete chemical substance record be centrally created and maintained. Each chemical substance record will consist of the following elements:

- A unique Chemical Abstracts Service (CAS) Registry number when one exists;
- A unique systematic name using the CAS 9th Collective Index naming convention when one exists;
- A unique EPA Chemical Registry name;
- A unique record number.

## III. Draft Implementation Procedures for the Chemical Identification Standard

A. Selection of the EPA Chemical Registry Name

The purpose of selecting an EPA Chemical Registry name is to provide a convenient to use name which will facilitate communication both internally and externally. This name will be selected with the following considerations:

- the name is correctly applied to the chemical substance
- the name is unambiguous
- the name is unique
- selection of the name takes into account the name(s) currently being used by EPA
- many chemicals will not require and EPA Chemical Registry Name, for these the CAS name will be sufficient

B. Use of Chemical Names

If an EPA Chemical Registry Name is chosen, it and the CAS 9th Collective Index name, will be the only names used by EPA. Implementation will require a period of transition to the new names. It is also recognized that certain circumstances may require exemption from the use of these names, exclusively. Synonyms will be available in the EPA Chemical Registry for query purposes.

C. Identification Numbers

The workgroup is currently divided on the use of the record number as part of this Standard. The positions of both sides are presented below.

1. Establishment of an EPA unique chemical identifier for all EPA chemical substances.
EPA needs a Chemical Identifier (ChemID) that is a simple, unique key applicable to all EPA chemicals, that is under our control, and that can be inserted readily into all Agency databases without necessitating significant changes in existing data. CAS Registry Numbers cannot be used exclusively for this purpose because they are not controlled by EPA, some EPA chemical substances will never have CAS numbers, and they have been routinely used incorrectly in Agency databases and documents.

CAS Registry Numbers are controlled by the Chemical Abstracts Services. CAS can, and regularly does, change CAS Registry Numbers whenever they decide to, without prior consultation or notification of anyone else. We cannot expect to create and maintain a stable and consistent Chemical Registry System (CRS), when our most critical data element is under outside control. Further, CAS and CAS alone decides which chemical entities should be assigned CAS Registry Numbers. Many chemical species of interest to EPA do not have CAS

Registry Numbers, and many never will. Another problem with using CAS REGISTRY NUMBERS so centrally is that it would lock the Agency into a permanent formal relationship with CAS, an independent, non-governmental entity. This would be unwise, both legally and financially.

The EPA Chemical Identification Standard is intended to provide a means of clearly and consistently identifying and distinguishing among all of the chemical species of interest to the Agency in all contexts where they may be referenced. In past years, Congress, in its environmental legislation, and the Agency, in its regulations and other formal communications, have used CAS REGISTRY NUMBERS as one means of identifying chemicals, often without conforming precisely to CAS's definitions (for example, use of the CAS Registry Number for a pure metal to identify the category of all chemicals containing that metal) . Also, some CAS REGISTRY NUMBERS that were correct in the past have since been changed. Consequently, there is a large legacy of incorrect CAS REGISTRY NUMBERS in EPA records and databases. Attempting at this point to use CAS REGISTRY NUMBER as the key identifier to link Agency data would require an across-the-board "clean-up" of the old CAS REGISTRY NUMBERS. Such an effort would be costly, would substantially delay implementation of the Standard, and would destroy potentially valuable historical data.

Any identifier that is part of the Chemical Identification Standard should be made available to EPA staff and others, including the public. If the Standard is implemented as planned, an EPA established identifier will be one of only two guaranteed unique identifiers of all EPA chemicals -- the other being the EPA Chemical Registry Name. Anyone wanting to distinguish unequivocally between two similar EPA chemical substances will need access to the ChemID as well as the EPA Name for verification. The ChemID will also be the only key linking chemical data in different EPA databases. Database managers and others wishing to connect electronically with EPA systems (other federal agencies, states, etc.) will need access to the ChemID to populate and maintain their own systems properly. In general, the ChemID may be more convenient to use than the EPA Name in some data management contexts. It will likely be shorter, more uniform, easier to sort by and index on (e.g., "EPA1234567" versus "Lead compounds -- SARA 313"). Finally, once in existence, the EPA ChemID will certainly be subject to requests under FOIA, along with the rest of the CRS data. So the question really is whether to create it, not whether it will be public.

2. Establishment of a unique chemical identifier only for EPA chemical substances without CAS Registry Numbers.

Only CAS Registry Numbers should be used to identify the majority of EPA chemical substances because they are the accepted identification standard for industry, government, and academia. Adding yet another identification number will only confuse people.

OPPT understands that chemicals having CAS Registry Numbers will have only these numbers as their public identification numbers and that EPA will attempt to obtain CAS Registry Numbers for chemicals as appropriate. Chemical substances without CAS numbers may need to display an internal EPA-specific ID number, until such time as a CAS number is available. Chemical categories or groups will need an EPA-specific ID number, and will need clear and explicit definitions of which chemicals are included in the category or group. In addition, if such numbers are to be used publicly, they should be intelligent, rather than unintelligent, ID numbers. (For example, use "EPA-CAT-######" as prefix to indicate that it is an EPA number for a category.)

## IV. Definitions

A. For the purposes of this Standard **Chemical Substances** were categorized and defined as follows:

A **Single Fully-defined** chemical substance consists of a single element or compound uniquely identified by the number and nature of the atoms present, the bonds which connect each of the atoms, the nature of the bonds, and the spatial orientation. With these descriptors known, there is no ambiguity as to the identity of the chemical. The nature of the atoms includes specification of atomic number, atomic weight for isotopes or tagged atoms (e.g., urea-14C, acetone-d6), and oxidation state (e.g., iron(II) chloride or ferrous chloride). The bonding and type of bond describes and differentiates straight and branched isomers (e.g., tert-butylacetic acid), bridgehead positions (e.g., bicyclo[2.2.1]hept-2-ene norbornene), substituent locations, and saturation/unsaturation and the associated location(s) (e.g., o-xylene or 1,2-xylene; methylacetamide; 1-octene; 1(3H)-isobenzofuranone). The spatial orientation unambiguously identifies structural or stereoisomers (e.g., cis-1,2-dichloroethene or (Z)-1,2-dichloroethene; (R)-2-chlorobutane or d-2-chlorobutane).

A chemical substance of **Known Composition** comprises two or more single fully-defined chemical substances that are always present in the same precisely-defined ratio(s). E.g., formulated mixtures, fixed-ratio salts, fully-defined isomer mixtures.

A chemical substance of **Variable Composition** is one for which all possible constituents are known single fully-defined chemical substances, but it is not known which constituent(s) is/are present or, if present, in what ratio(s). E.g., variable isomer mixtures, undefined-ratio salts, PCB mixtures (Aroclors).

An **Unknown Composition** chemical substance is one for which one or more constituents is/are unknown. This type of chemical substance may be described in terms of the original constituents of a reaction. E.g., complex reaction products, naturally occurring substances, biologicals.

A **Regulatory Chemical Class** is a set of individual substances grouped together due to chemical similarity for regulatory purposes. E.g., lead and lead compounds

A **Regulatory Category** is a set of individual substances grouped together due to similarity other than chemical for regulatory purposes. E.g., carpet adhesives.

A **Generic-CBI** chemical substance is one for which the Agency is not permitted to release an exact identity due to Confidential Business Information regulations.

## V. How to Comply with the Chemical Identification Standard

## VI. Obtaining Information about Chemical Substances

EPA will maintain a central Chemical Registry that will contain verified information identifying each chemical substance. The Chemical Registry will be available on the Internet. EPA offices may also

obtain all or portions of the Chemical Registry to incorporate mandatory attributes into their databases.

Any EPA office may request that a new chemical substance be added to the Chemical Registry. The office establishing the new chemical substance must apply to the data steward with the proposed EPA Chemical Name, the CAS number, CAS 9th Collective Index name and other mandatory attributes depending upon the type of chemical substance (see appendix, Table 1). If CAS identification is not provided the office must provide a justification for excluding it. The data steward, in conjunction with the Chemical Identification Standard Workgroup (see Roles and Responsibilities), will research the application and determine whether or not the chemical substance is already represented in the Chemical Registry and whether the attributes are correct.

## VII. Roles and Responsibilities

Implementation of the Chemical Identification Standard will result in a uniform, master chemical vocabulary for the Agency. Exclusive use of this precise vocabulary whenever the Agency or its representatives speak formally about chemicals will ensure that the regulated community, the public, and EPA itself, will always know the reference for each chemical. Developing and maintaining this vocabulary and ensuring adherence to the Chemical Identification Standard is an Agency-wide responsibility; however, organizationally one Office must assume a lead role. The lead Office must be determined before the Standard is implemented. Each Program Office that is not in a lead role will be responsible for participating in the development and maintenance of the master chemical vocabulary. With this in mind, it is recommended that a permanent staff be created to develop and maintain the chemical vocabulary. This staff would lead a permanent Chemical Identification Standard Standing Committee consisting of participants from all EPA Offices that routinely deal with chemicals.

The purpose of the Chemical Identification Standard Workgroup is to:

- Develop the initial master vocabulary, according to the Chemical Identification Standard including reconciling existing Agency information;
- Maintain the vocabulary as new information emerges;
- Ensure the consistency and quality of the vocabulary; and
- Resolve identification conflicts.

The Chief Information Officer (CIO) will lead the effort to build a system to house the master chemical vocabulary and will provide technical assistance to Program Offices who must include this data standard in their new or re-engineered systems. The CIO will also establish an agency process and operating procedures to review agency documents and publications for compliance with this standard.

Senior Information Resource Management Officers (SIRMOs) and Regional Information Resource Management (IRM) Branch Chiefs shall be responsible for assuring compliance with this standard within their information management environments.

The Office of Reinvention will coordinate between EPA's Programs and Regions and the Agency's State partners and stakeholders for the implementation of standards and business practices in EPA's national systems.

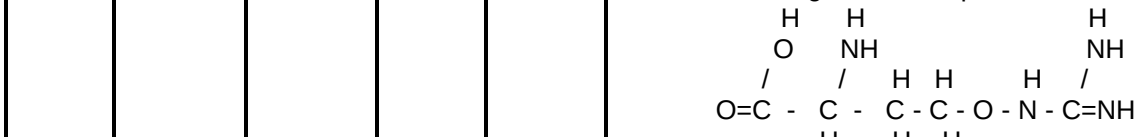
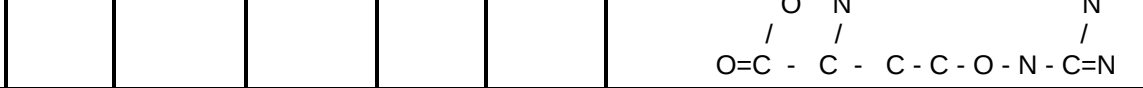
Each Program Office will be responsible for ensuring that new regulations, documents and automated systems identify chemicals according to this Standard.

## VIII. Implementation Schedule

**Appendix I**

**Table 1** Attributes for EPA Chemical Substances

| Attributes for the EPA Chemical Substances | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Attribute | Optionality( see Table 2) | | | | | | | Unique in clmn | Description and Comments |
| | Single Fully-Defined | Variable Compo-sition | Unknown Compo-sition | Known Compo-sition | Regulatory Chemical Classes | Regulatory Chemical Categories | Generic (CBI) | | |
| EPA Chemical Registry ID number | M | M | M | M | M | M | M | X | The EPA Chemical Registry identification number is a system derived unique, permanent identifier for each chemical substance. The EPA Chemical Registry ID number will not be reused so that a history of changes can be maintained. |
| CAS number | M | CS | CS | CN | CN | CN | B | X | The Chemical Abstracts Service Registry number for this chemical. This field will contain real CAS numbers only, displayed right justified with hyphens and no leading zeros. The system should provide output options for other formats. Expired CAS numbers will be maintained as history. |
| EPA Chemical Registry name | M | M | M | M | M | M | M | X | EPA Chemical Registry name is selected by EPA as the preferred name. This name should be unique to this chemical, but also be convenient to use. Organized rules will be developed for selection of this name. |
| EPA Chemical Registry name source | M | M | M | M | M | M | M | | The source of the EPA Chemical Registry name that contains enough information to find a reference. Organized rules will be developed for selection of a source. All sources need not be listed. |
| EPA Chemical Registry name context | CN | CN | CN | CN | CN | CN | CN | | The history, applicability, and, optionally, other industry(s) that use this name. A name can have more than one context. |

| Attributes for the EPA Chemical Substances | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Attribute | Optionality( see Table 2) | | | | | | | Unique in clmn | Description and Comments |
| | Single Fully-Defined | Variable Compo-sition | Unknown Compo-sition | Known Compo-sition | Regulatory Chemical Classes | Regulatory Chemical Categories | Generic (CBI) | | |
| molecular formula | M | CN | CN | CN | O | O | B | | The text that displays the number of atoms of each element in a molecule of a chemical substance. A chemical formula that indicates the kinds of atoms and the number of each kind in a molecule of a compound. Examples -- C6H12O6; C5H12N4O3 |
| structural formula<br><br>(linear structural formula) | CN | CN | CN | CN | O | O | B | | The text that displays the connectivity of atoms in a molecule of a chemical substance as a linear formula, such as SMILES. A chemical formula showing the linkage of the atoms in a molecule diagrammatically. Examples: H-O-H; O=C(O)C(N)CCONC(=N)N. |
| chemical structure<br><br>(graphical structural diagram) | CN | CN | CN | CN | O | O | B | | A graphical representation of a molecule of a chemical substance as a two or three-dimensional diagram. Example:<br>  <br>Note that H atoms can be omitted from a graphical representation, the same as in a structural formula, as in the following examples of the same substance:<br>  |
| molecular/atomic wt (formula wt) | M | CN (range) | O (range) | CN | O | O | B | | The sum of the atomic weights of constituent atoms in a molecule of a chemical substance. Example: 176.17 |

| Attributes for the EPA Chemical Substances | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Attribute | Optionality( see Table 2) | | | | | | | Unique in clmn | Description and Comments |
| | Single Fully-Defined | Variable Compo-sition | Unknown Compo-sition | Known Compo-sition | Regulatory Chemical Classes | Regulatory Chemical Categories | Generic (CBI) | | |
| systematic name | M | CS | CS | CS | CN | CN | B | X | A name that describes the chemical structure. When the chemical has a CAS Registry number, the CAS 9th Collective Index name will be used as the systematic name. The IUPAC name is the next best name and should be strongly encouraged due to international usage of the database. Prioritization rules will be established to determine which systematic name should be used if no CAS 9th Collective Index name is available. |
| administrative fields | TBD | TBD | TBD | TBD | TBD | TBD | TBD | | Who last updated; Error history field; original creator; creation date; QA/QC; fields for authorization; system version control; fields for review; others may be required |
| comment | O | O | O | O | O | O | M | | Used as a clarification of chemical identification. In addition to a general comment field, comment fields may be appropriate for a number of other fields such as name, systematic name, structure. |
| classification | CN | CN | CN | CN | CN | CN | CN | | Classification of a chemical according to its chemical structure, e.g. organophosphate, triazine. |
| chemical type | M | M | M | M | M | M | M | | chemical type indicator (single fully-defined, variable composition, etc.) |
| definition | O (B) | CIN | CIN | CIN | M | M | O | | Further description of the chemical name |

| Attributes for the EPA Chemical Substances | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Attribute | Optionality( see Table 2) | | | | | | | Unique in clmn | Description and Comments |
| | Single Fully-Defined | Variable Compo-sition | Unknown Compo-sition | Known Compo-sition | Regulatory Chemical Classes | Regulatory Chemical Categories | Generic (CBI) | | |
| synonyms | CS EPA O (Non-EPA) | CS EPA O (Non-EPA) | CS EPA O (Non-EPA) | CS EPA O (Non-EPA) | CS EPA O (Non-EPA) | CS EPA O (Non-EPA) | M for accession /submission No. | | Synonyms are alternate names and alternate IDs which could include names used for existing EPA regulations, historic EPA names and IDs from regulations, current and historic names from other EPA systems, or names submitted to EPA by the regulated industry<br>The purpose of including synonyms in this vocabulary is as an aid for searching. Synonyms will allow finding alternate names for searches of other databases; matching chemical names to the correct chemical; and finding legacy regulatory names. |
| synonym status | M | M | M | M | M | M | M | | If a synonym is known to be incorrect, it must be labeled as such. Examples for this are EPA errors such as typos, incorrect CAS numbers, misspellings. Including these types of synonyms would be left to the discretion of the Program Offices. |
| | O | O | O | O | O | O | O | | A synonym identified as one that could refer to more than one chemical. This is left optional because of the difficulty in determining a status for many synonyms. |
| | O | O | O | O | O | O | O | | Identify a synonym as a unambiguous (one-one match) if this is known to be true. Examples: carbon tetrachloride; methylene chloride. This is left optional because of the difficulty in determining a status for many synonyms. |
| synonym display | O | O | O | O | O | O | O | | If the synonym list for a particular chemical substance is excessively long, certain of the most common synonyms may be flagged to display. |

| Attributes for the EPA Chemical Substances | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Attribute | Optionality( see Table 2) | | | | | | | Unique in clmn | Description and Comments |
| | Single Fully-Defined | Variable Compo-sition | Unknown Compo-sition | Known Compo-sition | Regulatory Chemical Classes | Regulatory Chemical Categories | Generic (CBI) | | |
| synonym source | CD (syn) | CD (syn) | CD (syn) | CD (syn) | CD (syn) | CD (syn) | M | | The source of the synonym name that contains enough information to find a reference. Organized rules will be developed for prioritizing the selection of a source. All sources need not be listed. An EPA synonym should include as sources all major official databases. Synonyms used by entities outside EPA will require only one source. |
| synonym context | CN | CN | CN | CN | CN | CN | CN | | The history, applicability, and, optionally, other industry(s) that use this name. A name can have more than one context. |

**Table 2** Optionality Definitions

| Codes | Optionality Business Rules: |
|---|---|
| M | Mandatory -- Must have this piece of information |
| CS | Conditionally Required -- EPA must seek the data for this field |
| CIN | Conditionally Required -- If necessary for clarification of identity |
| CN | Conditionally Required -- If EPA has the data, it must be entered. If it is not available EPA does not have to seek it out. |
| CD | Conditionally Required -- Dependent on another Attribute. |
| CO | OR Conditional -- Must have one of the group of Attributes |
| O | Optional, Not Required -- Left to judgement of Data Manager. Include if practical or useful to provide it |
| B | Mandatory Blank |